# Native language detection with 'cheap' learner corpora

Julian Brooke and Graeme Hirst
University of Toronto

**Abstract**

We begin by showing that the best publicly available, multiple-L1 learner corpus, the *International Corpus of Learner English* (Granger *et al.* 2009), has issues when used directly for the task of native language detection (NLD). The topic biases in the corpus are a confounding factor that results in cross-validated performance that appears misleadingly high, for all the feature types which are traditionally used. Our approach here is to look for other, cheap ways to get training data for NLD. To that end, we present the web-scraped Lang-8 learner corpus, and show that it is useful for the task, particularly if large quantities of data are used. This also seems to facilitate the use of lexical features, which have been previously avoided. We also investigate ways to do NLD that do not involve having learner corpora at all, including double-translation and extracting information from L1 corpora directly. All of these avenues are shown to be promising.

**Keywords**: Native language, text classification, natural language processing, learner corpora

## 1. Introduction

Native language detection (or identification), henceforth NLD, is the task of distinguishing the native language background (L1) of a non-native writer of a text. As a natural language processing (NLP) task, it is properly categorized as a form of text classification, the standard approach to which is machine learning classification using algorithms such as support vector machines (SVMs) (Witten & Frank 2005). Generally speaking, these algorithms learn associations between features and classifications from a corpus of texts whose classification is known, and then use that information to classify new texts. Therefore, having a (preferably large) corpus of training data is a necessary first step for any machine learning approach. For NLD, there is a paucity of training (and testing) corpora, since roughly comparable texts from multiple L1 backgrounds (for the same L2) are required. The only three text corpora, as far as we are aware, that are appropriate for use with multiple language NLD are the *International Corpus of Learner English* (Granger *et al.* 2009) or ICLE, the *Cambridge Learner Corpus* (Yannakoudakis *et al.* 2011), and the *Falko Corpus* (Lüdeling *et al.* 2008). Most work in NLD has been carried out in the first of these, since the second is not publicly available (except for a very small portion that has recently been released) and the third is smaller and in German (the other two are English corpora). When only one

corpus is available, evaluation of NLD is often carried out using cross-validation, which involves building multiple models, training on one portion of the corpus and testing on another.

Koppel *et al.* (2005) was the first work in NLD to use the ICLE (version 1) (Granger *et al.* 2002); they trained SVM models with a set of stylistic features, including Part of Speech (POS) and character *n*-grams (sequences), function words, and spelling error types, achieving 80% accuarcy in a 5-language task. Tsur & Rappoport (2007) focused on character *n*-grams; they were concerned about the effect of topic bias, and attempted to remove it by discarding prominent words, with 66% accuracy in the same set as Koppel *et al.* (2005). Wong & Dras (2009), working in ICLE v2 (Granger *et al.* 2009), investigated particular kinds of syntactic error common to L1 backgrounds, but they failed to improve on Koppel *et al.* (2005). In recent work, Wong & Dras (2011) showed that generally syntactic patterns, as derived by a parser, are more effective than other stylistic features. The Cambridge Learner Corpus has been used recently by Kochmar (2011), who concluded that character *n*-grams are the most promising feature type, and Golcher & Reznicek (2011) used the Falko Corpus to show that title (topic) classification and NLD are more closely linked than previously assumed.

In Section 2 of this paper, we will begin by showing that the ICLE is problematic as a corpus for NLD because of topic bias in the corpus. Previous work had assumed that some features were immune from this problem, but we show conclusively that this is not the case. This motivates the primary goal of our work, which is the search for 'cheap' (easy to collect) alternatives. In Section 3, we introduce the Lang-8 learner corpus, which has been scraped from a website where language learners write a journal to improve their English. We show that this corpus can be used for NLD, but our results suggest that the task is much more difficult than suggested by previous work, and requires more data. In Section 4, we test two options for artificial learner corpora: doubly-translated texts from the LOCNESS corpus and *n*-gram information gleaned directly from L1 texts. In Section 5, we offer our conclusions.

# 2. Topic bias in the ICLE

## 2.1. Quantitative analysis

The ICLE, which in its current version contains 6,085 essays from 16 different languages, is intended to reflect, among other things, the state of EFL teaching in each of the countries around the world. An obvious challenge in building a corpus like the ICLE is incorporating the work of many researchers, educators and, of course, learners from different countries into a coherent whole. An original list of topics was chosen by the coordinating team, but leeway was clearly given to the organizers in each country, since some of the topics were, for instance, only relevant to Europeans (e.g. the future of a united Europe). Even when the original topic list was used, there were obvious biases in the particular topics chosen, with certain L1 backgrounds being dominated by certain topics (Granger *et al.* 2009: 6-7). This explains why many NLD researchers have avoided word features when working with the ICLE; a classifier can simply learn to distinguish L1 by distinguishing topics. However, the problem extends deeper than that: we believe that certain topics are correlated with entirely different registers,

which might have an effect on features that go beyond topic words. For example, many of the most common topics in the French subset of the corpus involve the relatively esoteric subjects of literature, religion, and politics, which might be discussed in a fairly formal register. In the Japanese corpus, however, we found a number of topics that were far more personal, for instance experience as an English learner and favorite travel destinations, which would likely be expressed in a more narrative and more colloquial manner. Arguably, these might reflect real differences in culture, but in the context of a corpus that cannot possibly reflect the full range of genres, we believe that these variations are extremely confounding for machine-learning based NLD, and they can affect a full range of feature types.

### 2.2. Experiment #1: Measuring the influence of topic

In order to test whether the variation in the ICLE is having an effect by artificially boosting classification accuracy, we carried out the first in a series of experiments. Most of our experiments in this paper will use the same feature set, which is a core set of features from previous work: function words, character *n*-grams, and POS *n*-grams; we also include word *n*-grams, which have mostly been avoided. Here *n*-grams include unigrams (single elements) and bigrams (pairs of elements). The POS is automatically tagged (Schmid 1995). Our input to the machine classifier consists of the normalized (to 1,000 words) frequency of these elements in the texts. For classification, we use the SVM algorithm included in the WEKA machine learning software (Witten & Frank 2005), with default settings. We report accuracy as the total number of texts whose author's L1 was correctly identified. Another constant is our notation for statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, though in different tables the numbers which are being compared may vary.

Our first experiment uses four languages from the ICLE: French, Spanish, Chinese, and Japanese. Using the annotation included in the ICLE, we took 200 texts from each sub-corpus and divided them into groups of 100 texts based on topic; all texts of a particular topic from the same L1 appeared in the same set. Then we took the same 200 texts, and just randomly distributed them into two sets. For each of these divisions, we trained a classifier on one set, and tested on the other, and then reversed the roles (so our results reflect the classification of all 800 texts). We hypothesized that if topic is playing a major role in NLD performance, the randomized version, which has topics which are spread across the training and testing sets, will be able to take advantage of that, while the topic-segmented version will not be able to.[1]

The results in Table 1 show that topic is indeed playing a major role in NLD, and the effect is not limited to words. All of the classifiers trained and tested using the topic-split version performed worse, and these results are highly significant. Character *n*-grams are by far the biggest offender; we posit that the distinct characters in words in the title are driving the high performance, and without that information character *n*-grams are simply not competitive. By contrast, the drop in word *n*-grams is fairly modest. But what is particularly troubling is the drop for POS *n*-grams and function words; these were supposed to be immune from the effects of topic, but clearly are not. Thus, we need to reconsider the validity of training and testing in the same corpus where topic variation is so confounding, and look for alternatives.

[1] Or at least will have less opportunity. Some of the different topics have certain common meta-topics (*e.g.* politics) which could still be used by the classifier. We did not attempt to control for this.

| Feature set | Randomized split | Topic-based split |
|---|---|---|
| Character *n*-grams | 82.9 | 50.2*** |
| POS *n*-grams | 92.1 | 85.0*** |
| Function words | 85.2 | 77.1*** |
| Word *n*-grams | 96.2 | 82.5*** |
| Combined | 96.5 | 86.1*** |

*Table 1. NLD accuracy (%) in ICLE, 4-language, randomized vs. topic-based cross-validation*

### 2.3. An alternative: filtering the ICLE

Before we move on to our own approaches, we will briefly consider one alternative, which has been recently proposed by Jarvis & Paquot (2012): filtering the ICLE at both the text level and the *n*-gram level to produce an unbiased corpus. Jarvis & Paquot suggest removing all texts from learners from Chinese, Japanese, Turkish, and Tswana backgrounds, since these have considerable variation from the others in terms of both topic and competency. However, these four groups share another important distinction: They represent all the non-European L1s in the corpus. That means in order to minimize these confounding effects, we would have to limit ourselves entirely to European languages, an entirely unacceptable compromise, since the properties of language transfer within closely related languages is likely to be entirely different from those between families; for example, Europeans may struggle with spelling errors between numerous close cognates, but this is not an issue for a Chinese speaker, who must instead contend with various lexical bundles that are directly translated across European languages but have no exact equivalent in Chinese. For native language detection as a real world task, a full range of languages must be considered. More generally, controlling for competency is a complicated problem because distance between L1 and L2 is likely to be a huge determining factor in competency; it is very difficult to separate the two and, if the goal is to improve performance of an algorithm for NLD, it is not clear that learner proficiency should be controlled for at all.

Moreover, Jarvis & Paquot removed *n*-grams that appeared both in prompts as well as commonly in the learner texts. Though this certainly would help remove some of topic bias, the examples they provide demonstrate the limitations of this approach: from one text, they remove *society* and *prison*, but preserve other topical words such as *punish*, *criminal*, and *rehabilitate*, which are just as problematic. Presumably, one could push this further, removing more and more words, but we predict that this would almost immediately impinge on *true* L1 transfer features (for instance, preferring a close cognate), undermining the ultimate goal of NLD. This approach can certainly be applied to improve the reliability of relevant language-transfer research, which is Jarvis & Paquot's interest, but, again, if the ultimate goal of the research is developing

robust high-performing NLD systems, discarding L1s and key features is not, we believe, a good way to begin.

# 3. A new 'cheap' corpus: The Lang-8 learner corpus

The construction of the ICLE was a major project that took several years. Though it will likely continue to expand, we doubt it will be repeated in the near future. However, it is not sufficient for our needs; as we saw in the previous section, it is biased in a way that interferes with the reliability of our results. However, a huge amount of non-native language is being produced constantly around the world, every day, and some of it is on the World Wide Web, making it accessible to us (if we can find it). The data we derived from such 'cheap' sources will almost certainly have noise, but, for data-driven approaches, it can make up for that fact by being available in almost unlimited quantities.

## 3.1. A description of the Lang-8 learner corpus

The Lang-8 website[2] provides a means for language learners to practice by writing journal entries in the language they are studying, which in turn is corrected by native speakers of that language who visit the site. We extracted a large collection of journals from the site, including 154,702 entries, or 22 million words. The site is based in Japan, and so learners of East Asian origin are disproportionately represented;[3] however, among the entries in our corpus there are 65 different native languages included, with 14 of those languages having at least 1,000 entries. Compared to the numerous variables that are recorded in manually collected learner corpora such as the ICLE, the information we have about each entry is rather minimal: other than (self-reported) native language and target language, we have a (unique) user name and the time which the entry was posted, though we use neither in the investigations reported here. There is some additional information available in the user profiles (e.g. gender), but we did not collect this information.

The ICLE contains primarily argumentative essays. The Lang-8 journal entries, by contrast, tend to be short personal narratives, though there are many exceptions: some users post their homework assignments, or ask for explicit translation or correction of a particular phrase out of the context of a coherent discourse. Though we did not carry out a rigorous analysis, the overall quality of the Lang-8 entries, i.e. the English proficiency of the users, seems to be generally much lower than the ICLE texts (which are written by university students). Moreover, the Lang-8 texts, because they are written entirely at the discretion of the user, appear to be more error-avoiding (Corder 1974); for the most part, users stay in their comfort zones, an effect which we posit is amplified by the knowledge that their text may be critiqued by native speakers. On the other hand, there are (presumably) no limits on the time or other resources that users may use to create the entries, so some entries may represent a fairly major investment, including revisions.

[2] http://lang-8.com

[3] The token counts for the best-represented L1s in the Lang-8 corpus, in millions of tokens, are as follows: Japanese, 7.79; Chinese (both Mandarin and Cantonese), 5.66; Korean, 4.31; Russian, 1.00; Spanish, 0.52; French, 0.39; German, 0.26; Polish, 0.25; Italian, 0.23; Vietnamese, 0.20; Indonesian, 0.20; Arabic, 0.19; Portuguese, 0.16; Thai, 0.15. All other L1s have less than 100,000 tokens.

## 3.2. Experiments

### *3.2.1. Experiment #2: Train on one, test on the other*

Armed with this new corpus, we wanted to see whether it could be used for NLD research, either by itself, or in concert with the ICLE. Our second experiment is a 7-language task (the 7 L1s, Spanish, French, Italian, Russian, Polish, Japanese, and Chinese, were selected because they are well represented in both corpora) with 200 texts per language per corpus. Since the texts in the Lang-8 corpus tend to be much shorter, our experimental texts actually consist of multiple original corpus texts; we combined them such that the average text lengths for the two corpora were comparable (about 500 words per text). Since we applied this same technique irrespective of languages, the text length across languages in the version of the Lang-8 that was used in the experiment is roughly uniform. The conditions for the experiment are: single corpus classification (10-fold cross-validation) with each of the two corpora, and training on one corpus and testing on another (that is, cross-corpus testing). The random baseline for this task is 14.2%, and our significance results reflect comparison to that baseline.

| **Feature set** | **ICLE, single corpus** | **ICLE, Lang-8 training** | **Lang-8, single corpus** | **Lang-8, ICLE training** |
|---|---|---|---|---|
| Character *n*-grams | 76.9*** | 22.6*** | 61.9*** | 22.0*** |
| POS *n*-grams | 83.8*** | 38.7*** | 70.0*** | 29.3*** |
| Function words | 70.1*** | 33.1*** | 60.2*** | 27.7*** |
| Word *n*-grams | 91.8*** | 44.6*** | 85.6*** | 27.6*** |
| Combined | 93.8*** | 46.1*** | 87.7*** | 26.7*** |

*Table 2. NLD accuracy (%), ICLE and Lang-8, 7-language, cross-validated vs. cross-corpus*

All of the results in Table 2 are well above chance. The best result, of course, is cross-validation in the ICLE, which is consistently higher than cross-validation in the Lang-8, though the conditions are similar. There are two explanations for this: the Lang-8 is a corpus with less distinction between L1s; or the biases in the ICLE that allow for easy classification by topic. Importantly, performance in the Lang-8 using the ICLE as training is extremely poor, suggesting that the patterns the classifier has learned in the ICLE do not generalize. In fact, they confound; we noticed that most Lang-8 texts are classified as Japanese, which, of the languages in the ICLE, seems to involve the most personal narratives (see Section 2.1). The other cross-corpus results, Lang-8 training and ICLE testing, are markedly better (though still well below the cross-validated), suggesting that the Lang-8 is a much better training corpus. Looking at the features,

we note that word *n*-grams tend to do the best, though POS *n*-grams are also useful. In any case, this task appears to be more difficult than cross-validated results in the ICLE would indicate.

### *3.2.2. Experiment #3: The more data, the better?*

One advantage of cheap data is the potential to get a lot of it. In the case of the Lang-8 corpus, the amount of data available for the European languages is actually much less than the ICLE, but the data available for two Asian languages, Japanese and Chinese, is much greater. Our next experiment tests the effect of adding more data. Does it improve the results we saw in Experiment #2? We created three Lang-8 training sets containing increasing amounts of Japanese and Chinese data, trained classifiers, and tested in the ICLE (200 texts for each language). The guessing baseline is 50%, but in Table 3, the significance testing reflects a comparison with the number in the previous column (i.e. did adding the data help?).

| **Feature set** | **200 text Lang-8** | **1,000 text Lang-8** | **5,000 text Lang-8** |
|---|---|---|---|
| Character *n*-grams | 39.3 | 68.5*** | 72.8 |
| POS *n*-grams | 63.8 | 80.0*** | 83.0 |
| Function words | 72.0 | 64.0 | 70.8** |
| Word *n*-grams | 69.8 | 83.5*** | 90.0*** |
| Combined | 59.8 | 80.0*** | 89.8*** |

*Table 3. NLD accuracy (%), train in Lang-8, test in ICLE, 2-language task, increasing data*

The result is clear: adding data makes a big difference. For all the features except function words (which are erratic in their performance), there is a statistically significant increase in performance. The pattern here is interesting: POS and character *n*-grams get a big boost when moving from 200 texts to 1,000 texts, but after that the increase is negligible. Word *n*-grams, however, get a major boost from both the increase from 200 to 1,000 and the increase from 1,000 to 5,000. This suggests that words, despite being mostly ignored or avoided in previous work, have the best long-term potential to improve NLD accuracy. Lexical choice almost certainly plays a key role in the phenomenon of language transfer (Odlin 1989), but we need a lot of data to properly identify the lexical items that are relevant, since there are many more of them than POS and function words.

# 4. Artificial learner corpora

Another approach to the problem of data scarcity in NLD is to side-step learner corpora altogether. This idea explicitly relies on the theory of language transfer (Odlin 1989), that the patterns in the L2 that distinguish the learners are a direct result of their L1. If this is the case, can the L1 itself be a source of information?

## 4.1. Experiment #4: Training on doubly-translated texts

Our first approach involves leveraging an automatic machine translation system to imbue English text with the style of a foreign language by translating through that language. For this purpose, we tested two web-based translation systems, Google Translate[4] and Yahoo! Babel Fish[5]. Starting with texts from the LOCNESS (Louvain Corpus of Native English Essays)[6], the native English counterpart from the ICLE, we translated them into each of four languages (again, Chinese, Japanese, French, and Spanish), and then translated them back into English. We used the result as a training set, with the version of the corpus translated through, for instance, Chinese, taking the place of a corpus of texts written by learners whose L1 is Chinese. One major advantage of this approach is that, since a single English corpus is used as the starting point, there is no possibility of topic bias. We test with both the ICLE and Lang-8 (200 texts for each language). The significance results are with respect to a comparison with the baseline, 25%.

| **Feature set** | **ICLE, BF** | **ICLE, Google** | **Lang-8, BF** | **Lang-8, Google** |
|---|---|---|---|---|
| Character *n*-grams | 29.6** | 31.6*** | 30.8*** | 32.0*** |
| POS *n*-grams | 27.0 | 38.0*** | 26.2 | 30.0** |
| Function words | 29.1** | 37.8*** | 21.8 | 31.4*** |
| Word *n*-grams | 30.5*** | 38.4*** | 25.9 | 34.7*** |
| Combined | 30.4*** | 38.8*** | 24.3 | 32.0*** |

*Table 4. NLD accuracy (%), train with LOCNESS doubly-translated texts, test in ICLE and Lang8, 4-language task*

[4] http://translate.google.com/
[5] http://babelfish.yahoo.com/
[6] http://www.uclouvain.be/en-cecl-locness.html

The results in Table 4 are not impressive, but, for those involving either Google or the ICLE, they are significantly above chance. Google Translate, which is known to be a statistical translation system, seems to reflect the L1 more than Babel Fish, which is thought to be rule-based. The discrepancy between ICLE and Lang-8 suggests that ICLE, independent of the bias, may also just be an easier test set for this task.

### 4.2. Experiment #5: NLD directly with L1 corpora

We briefly mention some other current work of ours that is also based on language transfer. Here, the idea is to extract information, particularly information related to lexical use, from the L1 corpus directly. We do not present the details here, but the basic idea is this: given a large corpus of some (non-English) L1, our software passes through the text, translating individual words and pairs of words into English using a bilingual lexicon, creating a database of counts. If a word or expression is commonly used in some L1, the direct English translation will have a high count. Then, if a particular word or expression appears in some non-native English text, we can look up the counts in our database, creating a ratio for each pair of languages. We sum the ratios across all the uncommon words and (two-word) expressions in the text to get a set of total ratios for the texts, and then sum across languages to get a score for each language; if a text has a high Chinese score relative to other texts, it will be classified as Chinese. Table 5 shows some preliminary results for Chinese and Japanese, with 'training' (collecting counts) in 100 million blog corpora (Burton *et al.* 2009).

| **Feature set** | **ICLE** | **Lang-8** |
|---|---|---|
| Word *n*-grams | 67.3*** | 66.0*** |

*Table 5. NLD accuracy (%), L1 method, test in ICLE and Lang-8*

For this method, we only use *n*-grams (POS features do not translate well). The performance in Table 5 is fairly good; for the ICLE, it is similar to the results when training with 200 texts from the Lang-8. There are two key advantages of this method: first, L1 corpora are usually available in abundance. Second, it allows for us to build a correspondence between forms in the L2 and their source forms in the L1; this could be useful for applications such as automated error correction (Leacock *et al.* 2010).

## 5. Conclusion

In this paper we have argued that the mostly commonly used multiple-L1 learner corpus, the ICLE, has problems when applied as is to the task of native language detection; in particular, the topic biases resulting from the way the corpus was built is a confounding factor that results in cross-validated performance that is misleading, and training that results in near chance performance elsewhere. Our approach here was to look for other, cheap ways to get training data for NLD; we presented the web-scraped Lang-8 learner corpus, and showed that it is useful for the task, particularly if large

quantities of data are used. We also investigated ways to do NLD that do not involve having learner corpora at all. All of these avenues are promising, and future work will include seeing how they might be combined to create a state-of-the-art system.

## References

Burton, K., Java, A. & Soboroff, I. (2009). The ICWSM 2009 Spinn3r Dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media* (ICWSM '09), San Jose, CA.

Corder, S.P. (1974). Error analysis. In Allen J.B.P. & Corder, S.P (eds) *The Edinburgh Course in Applied Linguistics. Volume 3 – Techniques in Applied Linguistics*. Oxford: Oxford University Press, 122-131.

Golcher, F. & Reznicek, M. (2011). Stylometry and the interplay of title and L1 in the different annotation layers in the Falko corpus. In *Proceedings of Quantitative Investigations in Theoretical Linguistics 4*. Berlin, 39-34.

Granger S., Dagneaux E. & Meunier F. (2002). *The International Corpus of Learner English.* Handbook and CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S., Dagneaux E., Meunier, F. & Paquot, M. (2009). *International Corpus of Learner English (Version 2)*. Louvain-la-Neuve: Presses universitaires de Louvain.

Jarvis, S. & Paquot M. (2012). Exploring the role of *n*-grams in L1 identification. In Jarvis S. & Crossley S. A. (eds.) *Approaching Language Transfer through Text Classification. Explorations in the Detection-based Approach*. Multilingual Matters: Bristol, 71-105.

Kochmar, E. (2011). *Identification of a Writer's Native Language by Error Analysis*. Master's thesis, University of Cambridge.

Koppel, M., Schler, J. & Zigdon, K. (2005). Determining an author's native language by mining a text for errors. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (KDD '05), 624-628.

Leacock, C., Chodorow, M., Gamon, M. & Tetreault, J. (2010). *Automated Grammatical Error Detection for Language Learners*. Morgan & Claypool.

Lüdeling, A., Doolittle, S., Hirschmann, H., Schmidt, K. & Walter, M. (2008). Das Lernerkorpus Falko. *Deutsch als Fremdsprache* 42, 67-73.

Odlin, T. (1989). *Language Transfer. Cross-Linguistic Influence in Language Learning*. Cambridge: Cambridge University Press.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT Workshop,* 47-50.

Tsur, O. & Rappoport, A. (2007). Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition* (CACLA '07), 9-16.

Witten, I. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. San Diego, CA: Morgan Kaufmann.

Wong, S.J. & Dras, M. (2009). Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop*, 53-61.

Wong, S.J. & Dras, M. (2011). Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1600-1610.

Yannakoudakis, H., Briscoe, T. & Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 180-189.